Biometrics 62, 28–35
March 2006

DOI: 10.1111/j.1541-0420.2005.00406.x

Haplotype-Based Association Analysis in Cohort and Nested Case–Control Studies

Jinbo Chen* **and Nilanjan Chatterjee**

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, Rockville, Maryland 20852, U.S.A.
**email:* chenjin@mail.nih.gov

Summary. Genetic epidemiologic studies often collect genotype data at multiple loci within a genomic region of interest from a sample of unrelated individuals. One popular method for analyzing such data is to assess whether haplotypes, i.e., the arrangements of alleles along individual chromosomes, are associated with the disease phenotype or not. For many study subjects, however, the exact haplotype configuration on the pair of homologous chromosomes cannot be derived with certainty from the available locus-specific genotype data (phase ambiguity). In this article, we consider estimating haplotype-specific association parameters in the Cox proportional hazards model, using genotype, environmental exposure, and the disease endpoint data collected from cohort or nested case–control studies. We study alternative Expectation-Maximization algorithms for estimating haplotype frequencies from cohort and nested case–control studies. Based on a hazard function of the disease derived from the observed genotype data, we then propose a semiparametric method for joint estimation of relative-risk parameters and the cumulative baseline hazard function. The method is greatly simplified under a rare disease assumption, for which an asymptotic variance estimator is also proposed. The performance of the proposed estimators is assessed via simulation studies. An application of the proposed method is presented, using data from the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study.

Key words: Cohort study; Cox proportional hazards model; Nested case–control study; Unphased genotype data.

1. Introduction

Genetic association studies based on marker genotype data collected from a sample of unrelated individuals are now widely used to study genetic mechanisms for complex diseases. Haplotype-based association analysis, that is, the study of the association between a disease phenotype and the combinations of alleles at multiple loci along individual chromosomes, has been argued to be more powerful than single-locus analysis for detecting gene–disease association (Risch and Merikangas, 1996; Botstein and Risch, 2003). However, current molecular technology for directly identifying haplotypes is expensive and is not feasible for large-scale epidemiologic studies. Instead, typical studies collect locus-specific genotype data that may not be fully informative regarding which set of alleles lies on a particular chromosome (phase ambiguity). For example, if A/a and B/b denote the major/minor alleles in two di-allelic loci, then subjects with genotypes (Aa) and (Bb) at the first and second locus, respectively, are considered "phase ambiguous": their genotypes could arise from either the haplotype pair ($A - B$, $a - b$) or the haplotype pair ($A - b$, $a - B$). Uncertainty in phase could also arise due to missing genotype information for some loci. Thus, assessing disease-haplotype association with only locus-specific genotype data is a special missing data problem.

In recent years, various researchers have developed a number of methods for detecting haplotype-based associations in the presence of phase ambiguity, using the case–control epidemiologic study design (Fallin and Schork, 2000; Schaid et al., 2002; Epstein and Satten, 2003; Stram et al., 2003; Zhao, Li, and Khalid, 2003). In this article, we propose methods for haplotype analysis for cohort and nested case–control studies, two widely used prospective epidemiologic study designs. Cohort studies collect biologic samples and questionnaire data at baseline on a group of healthy subjects, who are subsequently followed over a period of time to record information on disease incidence and age at onset/censoring. For rare diseases such as cancer, nested case–control studies (Thomas, 1977) conducted within an established cohort are now increasingly being used for genetic association studies. Studies using this design limit the genotyping effort and collection of expensive biomarker information only to cases and a small number of controls matched on follow-up time of the cases.

Interest has thus arisen in assessing the association between disease phenotypes and haplotypes using data collected from cohort and nested case–control studies. Lin (2004) recently proposed a nonparametric maximum-likelihood (NPMLE) method for testing and estimation of haplotype effects in the Cox proportional hazards (CPH) model using data from cohort studies. In this article, we develop an alternative computationally simple method for estimating haplotype-specific risk parameters in the CPH model, with data available from either cohort or nested case–control studies.

In Section 2, we present the model and notation. In Section 3, we propose alternative methods for unbiased estimation of haplotype frequencies in the underlying population. In Section 4, we present methods for estimating the relative-risk and the baseline hazard parameters of the CPH model. In the same section, we show how a rare-disease approximation leads to a remarkably simple way of estimating the relative-risk parameters and an associated asymptotic variance–covariance matrix. In Section 5, the methods are illustrated using data from the Alpha-Tocopherol, Beta-Carotene (ATBC) Cancer Prevention Study (Woodson et al., 2003). In Section 6, we study the finite-sample properties of the estimator, using simulated nested case–control studies involving four marker loci in the genomic region GPX1. In this section, we also evaluate the efficiency of the proposed method relative to the NPMLE approach of Lin (2004) for analysis of cohort studies. The article concludes in Section 7 by summarizing some of the computational and practical advantages of the proposed method.

2. Data Structure and Model Specification

Throughout this article, we assume, without loss of generality, that the underlying time scale of disease incidence is biologic age. Let T denote the true age at onset of the disease and C denote the censoring time. The observed phenotype for an individual can be represented as $[\Delta = I(T \leq C),\ X = \min(T,\ C)]$, where I is the indicator function. In full-cohort studies, the multilocus genotype information $\mathbf{G}$ and data on some possibly time-varying covariates $Z(t)$ are collected on all subjects. Suppose, K out of n subjects in the cohort develop disease during the study period at K distinct disease times $t_1 < t_2 < \cdots < t_K$. We define $\mathcal{R}_k$ to be the set of all subjects in the cohort at risk at t_k $(X \geq t_k)$, and let n_k be the number of subjects in $\mathcal{R}_k$. For nested case–control studies, $m - 1$ controls are sampled without replacement from the nondiseased subjects in $\mathcal{R}_k$, and the sample consists of all cases in the cohort and the sampled controls. Let $\tilde{R}_k$ be the subset of all controls sampled from $\mathcal{R}_k$ together with the kth case. The genotype information $\mathbf{G}$ and covariates $Z(t)$ are collected only for subjects in $\{\tilde{R}_k : k = 1, \ldots, K\}$.

For a given data set, let $\mathcal{H} = \{\ldots, h_r, \ldots, h_s, \ldots\}$ denote the set of all possible haplotypes, and let $D = (h_r,\ h_s)$ denote the two haplotypes (the diplotype) an individual carries in his/her pair of homologous chromosomes. In presence of phase ambiguity, a multilocus genotype $\mathbf{G}$ could be consistent with multiple diplotypes when $\mathbf{G}$ is heterozygous at two or more loci. We will denote $\mathcal{D}_{\mathbf{G}}$ to be the set of all possible diplotypes compatible with $\mathbf{G}$.

Traditionally, the CPH model is the method of choice for analyzing data from cohort and nested case–control studies. Similar to Lin (2004), we propose to quantify the haplotype-associated disease risk using the CPH model as follows. The disease hazard at age t for an individual with diplotype D and covariates $Z(t)$ is modeled as

$$\lambda[t \mid D, Z(t)] = \lambda_0(t) e^{\beta_D + \beta_z' Z(t)}, \tag{1}$$

where $\lambda_0(t)$ is a nonparametric baseline hazard function, and β_D is the relative hazard (relative risk) associated with diplotype D in reference to the baseline diplotype D_0. One may impose more structural assumptions on the risk associated with D (Wallenstein, Hodge, and Weston, 1998; Epstein and Satten, 2003; Zhao et al., 2003). For example, the following three models can be used: (i) $\beta_{(h_r,h_s)} = \beta_{h_r} + \beta_{h_s}$ (additive model), (ii) $\beta_{(h_r,h_s)} = I(h_r = h_s)\beta_{h_r} + I(h_r \neq h_s)(\beta_{h_r} + \beta_{h_s})$ (dominant model), and (iii) $\beta_{(h_r,h_s)} = I(h_r = h_s)\beta_{h_r}$ (recessive model). Let β_g denote the vector of regression parameters associated with all diplotypes/haplotypes. Let β denote $(\beta_g,\ \beta_z)$. We will describe all of the methods in the context of model (1), but note that the proposed methods can easily be extended to include, for example, the interaction terms.

3. Estimation of Haplotype Frequencies

As will be seen later, estimation of parameters in the CPH model (1), in the presence of phase ambiguity, requires knowing the frequencies of the different haplotypes in the underlying study population. In this section, we describe methods for estimating haplotype frequencies.

We assume that the population under study is in Hardy–Weinberg equilibrium (HWE). That is, if f_r denotes the frequency of the rth haplotype, the frequency of a diplotype $(h_r,\ h_s)$ is given by $2f_r f_s$ if $r \neq s$ and f_r^2 if $r = s$. Let $\mathbf{f}$ denote the vector of haplotype frequencies. Using unphased genotype data from a random sample of unrelated individuals, Excoffier and Slatkin (1995) proposed to apply an Expectation-Maximization (EM) algorithm for the estimation of haplotype frequencies. For full-cohort studies, where genotype data are available on all members, such an EM algorithm can be directly applied.

For nested case–control studies, however, cases and the matched controls who have been selected to be genotyped cannot be treated as a representative sample from the underlying population. We consider two options for estimating haplotype frequencies for this design. In one, we estimate the haplotype frequencies based on the EM algorithm applied to genotype data only from controls. Under the assumption that the disease is rare and that the censoring mechanism is unrelated to the genomic region under study, the controls can be treated as a representative sample from the underlying population. Thus the estimates of haplotype frequencies based on only controls would be approximately unbiased. In an alternative approach, we propose to use genotype data from all cases and controls in the nested case–control sample but to account for differential sampling of cases and controls using the Horvitz–Thompson (Horvitz and Thompson, 1952) weighting approach. Specifically, we propose to weight the contribution of each nonduplicated subject by the inverse of the sampling probability for that subject being included in the nested case–control sample. Typically, all cases from the underlying cohort would be selected, and thus the sampling probability for each case would be given by $\pi = 1$. Samuelsen (1997) showed that for a nested case–control design, the sampling probability for a nondiseased subject who is followed up to age X is given by

$$\pi = 1 - \prod_{k:t_k \leq X} \left(1 - \frac{m-1}{n_k - 1}\right).$$

Using these weights, we propose to use a weighted EM algorithm where estimates of haplotype frequencies are iteratively updated using the formula

$$f_r^{(s+1)} = \frac{\displaystyle\sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{D\in\mathcal{D}_{\mathbf{G_i}}} \frac{\mathrm{pr}_{\mathbf{f}^{(s)}}\left[D = (h_k, h_l)\right] v^r_{(kl)}}{\displaystyle\sum_{D'\in\mathcal{D}_{\mathbf{G_i}}} \mathrm{pr}_{\mathbf{f}^{(s)}}\left[D' = (h'_k, h'_l)\right]}}{2\displaystyle\sum_{i=1}^{n}\frac{1}{\pi_i}},$$

where $v^r_{(kl)} = 2$ if $k = l = r$, $v^r_{(kl)} = 0$ if $k \neq r$ and $l \neq r$, and $v^r_{(kl)} = 1$ otherwise. For full-cohort studies, where $\pi_i = 1$ for all subjects, the above algorithm reduces to the EM algorithm of Excoffier and Slatkin (1995).

4. Estimation of Relative-Risk Parameters β

Since only $\mathbf{G}$ is observed but not D, standard Cox regression analysis cannot be performed based on model (1). Following Prentice (1982), we derive the hazard function for disease conditional on the observable genotype data $\mathbf{G}$ and covariates $Z(t)$. Let $\tilde{Z}(t) = \{Z(s), 0 \leq s \leq t\}$ denote the history of $Z(t)$ up to time t. Let $\Lambda_0(t) = \int_0^t \lambda_0(t)\, dt$ denote the cumulative baseline hazard function. Based on the diplotype-specific hazard model given in equation (1), we derive the induced model for $\lambda[t \,|\, \mathbf{G}, \tilde{Z}(t)]$ in the form

$$\lambda[t \,|\, \mathbf{G}, \tilde{Z}(t)] = \lambda_0(t) e^{\beta'_z Z(t)} r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)], \qquad (2)$$

where

$$r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)] = \frac{\displaystyle\sum_{D\in\mathcal{D}_{\mathbf{G}}} e^{\beta_D}\mathrm{pr}[T > t \,|\, D, \tilde{Z}(t)]\mathrm{pr}_{\mathbf{f}}(D)}{\displaystyle\sum_{D\in\mathcal{D}_{\mathbf{G}}} \mathrm{pr}[T > t \,|\, D, \tilde{Z}(t)]\mathrm{pr}_{\mathbf{f}}(D)},$$

$$\mathrm{pr}[T > t \,|\, D, \tilde{Z}(t)] = \exp\left[-e^{\beta_D}\int_0^t e^{\beta'_z Z(s)} d\Lambda_0(s)\right],$$

and the diplotype frequencies $\mathrm{pr}_{\mathbf{f}}(D)$ are defined in terms of haplotype frequencies $\mathbf{f}$ assuming HWE (see Section 3). In equation (2), $e^{\beta'_z Z(t)} r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)]$ can be viewed as a generalized relative-risk function that describes the relative risk associated with the genotype $\mathbf{G}$ and covariates $\tilde{Z}(t)$ in reference to the baseline hazard $\lambda_0(t)$. Unlike the original CPH model (1), the induced model for genotype-specific hazard does not follow the proportional hazards form; the function $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)]$ depends on the time-dependent cumulative baseline hazard function $\Lambda_0(t)$.

4.1 *Estimation of β and $\Lambda_0(t)$ for Cohort Studies*

We propose jointly estimating β and $\Lambda_0(t)$ based on the induced hazard function $\lambda[t \,|\, \mathbf{G}, \tilde{Z}(t)]$, with the haplotype frequencies $\mathbf{f}$ fixed at their estimated value $\hat{\mathbf{f}}$ obtained using the EM algorithm described in the previous section. For estimation of β, we propose to use the partial likelihood function (Cox, 1972) based on the induced hazard model $\lambda[t \,|\, \mathbf{G}, \tilde{Z}(t)]$, the formula for which is given by

$$\mathrm{PL} = \prod_{k:\Delta_k=1} \frac{e^{\beta'_z Z_k(t_k)} r_{\mathbf{G}_k,\tilde{Z}_k(t_k)}[t_k; \mathbf{f}, \beta, \Lambda_0(\cdot)]}{\displaystyle\sum_{l\in R_k} e^{\beta'_z Z_l(t_k)} r_{\mathbf{G}_l,\tilde{Z}_l(t_k)}[t_k; \mathbf{f}, \beta, \Lambda_0(\cdot)]}. \qquad (3)$$

We observe that partial likelihood (PL) involves not only β but also the cumulative baseline hazard function $\Lambda_0(t)$. Motivated by Breslow's estimator (1972, 1974), we propose an estimating equation for $\Lambda_0(t)$ as

$$\Lambda_0(t_j) = \sum_{k:t_k\leq t_j} \frac{1}{\displaystyle\sum_{l\in R_k} e^{\beta'_z Z_l(t_k)} r_{\mathbf{G}_l,\tilde{Z}_l(t_k)}[t_k; \mathbf{f}, \beta, \Lambda_0(\cdot)]}. \qquad (4)$$

We now propose to estimate β and $\Lambda_0(t)$ by iterating the following steps:

1. Given current estimates $\hat{\Lambda}_0^{(s)}(t)$ and $\hat{\beta}^{(s)}$, we maximize the PL as a function of β_g with $\mathrm{pr}[T > t \,|\, D, \tilde{Z}(t)]$ fixed at $\mathrm{pr}_{[\hat{\Lambda}_0^{(s)}(t),\hat{\beta}^{(s)}]}[T > t \,|\, D, \tilde{Z}(t)]$ in the formula of $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)]$ (see equation (3)) and $e^{\beta'_z Z(t)}$ fixed at $e^{\hat{\beta}_z^{(s)'} Z(t)}$. This step updates $\hat{\beta}_g^{(s)}$ to $\hat{\beta}_g^{(s+1)}$.
2. Maximize the PL as a function of β_z with $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta_g, \beta_z, \Lambda_0(\cdot)]$ fixed at $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \hat{\beta}_g^{(s+1)}, \hat{\beta}_z^{(s)}, \hat{\Lambda}_0^{(s)}(\cdot)]$. This step can be performed by weighted Cox regression using existing software such as S-plus. Thus, $\hat{\beta}_z^{(s)}$ is updated to $\hat{\beta}_z^{(s+1)}$.
3. Obtain $\hat{\Lambda}_0^{(s+1)}(t)$ using the right-hand side of formula (4) with $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \beta, \Lambda_0(\cdot)]$ fixed at $r_{\mathbf{G},\tilde{Z}(t)}[t; \mathbf{f}, \hat{\beta}^{(s+1)}, \hat{\Lambda}_0^{(s)}(\cdot)]$.

The above algorithm takes advantage of the fact that the induced hazard function $\lambda[t \,|\, \mathbf{G}, \tilde{Z}(t)]$ retains the proportional hazards form in $Z(t)$. This is useful since direct maximization of the PL over a large number of parameters in the presence of time-dependent covariates can be numerically challenging. The theoretical derivation of the asymptotic variance of $\hat{\beta}$ is complex and would require modern semiparametric inference theory. We suggest obtaining the asymptotic variance of $\hat{\beta}$ using a nonparametric bootstrap sampling method that naturally takes into account the additional variation due to estimation of the nuisance parameters $\mathbf{f}$ and $\Lambda_0(t)$.

4.2 *Estimation of β and $\Lambda_0(t)$ for Nested Case–Control Studies*

The PL function for nested case–control studies takes the same form (equation (3)) as that for cohort studies (Oakes, 1981). The estimator for the cumulative baseline hazard $\Lambda_0(t)$, however, needs to be modified to account for the outcome-dependent sampling design. Motivated by Goldstein and Langholz (1992), we propose the estimating function

$$\Lambda_0(t_j) = \sum_{k:t_k\leq t_j} \frac{1}{\dfrac{n_k}{m}\displaystyle\sum_{l\in\tilde{R}_k} e^{\beta'_z Z_l(t_k)} r_{\mathbf{G}_l,\tilde{Z}_l(t_k)}[t_k; \mathbf{f}, \beta, \Lambda_0(\cdot)]}, \qquad (5)$$

where the weight n_k/m in the denominator takes into account the risk set sampling of the nested case–control design. The iterative algorithm described for the cohort study now applies with the modified estimator of $\Lambda_0(t)$.

4.3 *Estimation of β and $\Lambda_0(t)$ Assuming Rare Disease*

We consider a simplification of the estimation method, based on a rare-disease approximation, that applies to both cohort and nested case–control studies. The same approximation also leads to a simple asymptotic variance formula for the relative-risk parameters. When the disease is rare, one can assume $\mathrm{pr}[T > t \mid D, \tilde{Z}(t)] \approx 1$, so that

$$\lambda[t \mid \mathbf{G}, Z(t)] \approx \lambda_0(t) r_{\mathbf{G},Z(t)}(\mathbf{f}, \beta), \tag{6}$$

where

$$r_{\mathbf{G},Z(t)}(\mathbf{f}, \beta) = e^{\beta_z' Z(t)} \sum_{D \in \mathcal{D}_{\mathbf{G}}} e^{\beta_D} \mathrm{pr}_{\mathbf{f}}(D \mid \mathbf{G}).$$

Under this approximation, the relative-risk function $r_{\mathbf{G},Z(t)}(\mathbf{f}, \beta)$ does not depend on the baseline hazard function $\lambda_0(t)$. This leads to a remarkably simple way of estimating the relative-risk parameters β. As before, we assume that haplotype frequencies $\mathbf{f}$ have been estimated by $\hat{\mathbf{f}}$ using an appropriate EM algorithm. Then, β can be estimated by maximizing the PL

$$\mathrm{PL}^{\mathrm{rare}} = \prod_{k:\Delta_k=1} \frac{e^{\beta_z' Z_k(t_k)} \sum_{D \in \mathcal{D}_{\mathbf{G}_k}} e^{\beta_D} \mathrm{pr}_{\hat{\mathbf{f}}}(D \mid \mathbf{G}_k)}{\sum_{l \in \tilde{R}_k} e^{\beta_z' Z_l(t_k)} \sum_{D \in \mathcal{D}_{\mathbf{G}_l}} e^{\beta_D} \mathrm{pr}_{\hat{\mathbf{f}}}(D \mid \mathbf{G}_l)},$$

with respect to β. Further, the asymptotic variance of β can be estimated as

$$\hat{I}_{\beta\beta}^{-1}/n + \hat{I}_{\beta\beta}^{-1} \hat{I}_{\beta\mathbf{f}} \left(\hat{I}_{\mathbf{ff}}^{n_c}\right)^{-1} \hat{I}_{\beta\mathbf{f}}^{T} \hat{I}_{\beta\beta}^{-1}/n_c,$$

as derived in the Appendix.

5. Application to Data from the ATBC Study

We applied our methods to analysis of data from the ATBC Cancer Prevention Study conducted in Finland, a large, randomized cancer prevention trial of male smokers assessing the efficacy of supplementation with alpha-tocopherol, beta-carotene, or both in reducing the incidence of lung, prostate, and other cancers. The ATBC study cohort consisted of 29,133 white males who smoked at least five cigarettes daily. The trial began in 1985 and ended in 1993, and subjects continued to be followed afterward. A nested case–control study of prostate cancer was conducted within the cohort of ATBC subjects who provided a whole-blood sample. From January 1, 1983 to December 31, 1994, 208 incident cases were observed, and for each case a control was selected from the cohort matched on length of follow-up, age at randomization (±5 years), intervention group, and study clinic (Woodson et al., 2003). We excluded from the study 15 case–control pairs for both of which no single-nucleotide polymorphism (SNP) genotype data were available. It was of interest to assess whether the gene IL1A affected the risk of prostate cancer and whether it modified the efficacy of alpha-tocopherol treatment.

Two polymorphisms, IL1A889 (A/G) and IL1A4845 (T/C), were genotyped within the IL1A region. Out of the remaining 193 case–control pairs, 55 pairs had incomplete genotype data for IL1A polymorphisms. We estimated the haplotype frequencies based on controls to be 0.666/0.020/0.023/0.290 for the haplotypes AT/AC/GT/GC, respectively. The two polymorphisms were in strong linkage disequilibrium ($D' = 0.9$), and neither of the loci significantly departed from HWE. For illustrative purposes, we assumed an additive model for the haplotype effect and applied our methods adopting the rare-disease approximation. Setting AT as the reference haplotype, the log-relative-risk parameter estimates for haplotype GC, the rare-haplotype category AC/GT, and the interaction between GC and intervention were estimated to be −0.051, −0.312, and 0.205, respectively, with corresponding standard deviations (SDs) 0.230, 0.342, and 0.336. The analysis including only main effects for GC and AC/GT led to log-relative-risk parameter estimates (SDs) 0.043 (0.163) and −0.342 (0.391), respectively. Thus, the current investigation did not reveal any significant association between IL1A haplotypes and the risk of prostate cancer, nor did it show any significant modification of the intervention effect by IL1A haplotypes. The investigators are now accumulating more cases to increase the power of the study.

6. The Simulation Study

6.1 *Finite-Sample Performance*

We evaluated the performance of the proposed test using simulated genotype data in the context of GPX1. We selected four common SNPs based on the resequencing data from a project at the National Cancer Institute for 31 Caucasian American subjects. The haplotypes that we reconstructed were 1112, 1111, 1212, 2121, 1211, 1121, and 2211, with “1” referring to the wild-type and “2” to the variant allele. The corresponding frequencies were 0.298, 0.267, 0.152, 0.117, 0.099, 0.034, and 0.032, respectively.

We first generated the diplotype data for a cohort of 3000 subjects assuming HWE based on the haplotype frequencies above. We simulated the time-to-disease data using the exponential hazard model $\lambda[t \mid D = (h_r, h_s)] = \lambda_0 e^{\beta_{h_r} + \beta_{h_s}}$, assuming additive haplotype effects. The rare haplotypes 2121, 1211, 1121, and 2211 were combined into a single “rare haplotypes” group, and haplotype 1112 was used as the baseline. Thus, the disease-risk model involved three relative-risk (β) parameters. The random censoring time C was generated from the exponential distribution with hazard function $\gamma(t) = \gamma$. The parameters λ_0 and γ were chosen in such a way that approximately 10% of the subjects in the whole cohort would be cases ($\Delta = 1$). For each simulated cohort, a nested case–control sample was drawn using a one-to-one matching ratio. We then increased the amount of phase ambiguity by further deleting the genotype information for each individual marker for 20% of the subjects, randomly selected from the nested case–control sample. The simulation study was replicated 200 times.

Table 1 shows the mean and SD of the estimated haplotype frequencies for the simulated data. When data on both cases and controls were used, the weighted EM algorithm (W-EM) appeared to give consistent estimates, but when $\beta \neq 0$, the unweighted EM algorithm could produce substantial bias. Under the null hypothesis ($\beta = 0$), the unweighted EM algorithm was also consistent and produced more precise estimates. This is not surprising because, when the disease risk is not associated with haplotypes, the haplotype distribution in the nested case–control sample is still representative of that in the underlying population. Thus, in this case, the unweighted EM algorithm yields the efficient maximum-likelihood estimates. Estimation using the unweighted EM algorithm applied to only controls (EM rare) performed remarkably well, both in terms of efficiency and unbiasedness.

Table 1

*Simulated nested case–control studies involving the GPX*1 *gene: performance of different EM algorithms for estimation of* **f**. *Cohort size* = 3000, *control/case matching ratio* = 1, *number of cases* ≈ 300.

	β = (0, 0, 0)			β = (1.5, 0.4, 0)		
	W-EM (SD_E)[a]	EM (SD_E)[b]	EM-rare (SD_E)[c]	W-EM (SD_E)	EM (SD_E)	EM-rare (SD_E)
f_1	0.298 (0.041)	0.300 (0.018)	0.298 (0.024)	0.301 (0.040)	0.242 (0.014)	0.312 (0.021)
f_2	0.269 (0.042)	0.267 (0.017)	0.268 (0.023)	0.264 (0.034)	0.388 (0.016)	0.239 (0.019)
f_3	0.156 (0.033)	0.151 (0.015)	0.153 (0.019)	0.152 (0.030)	0.140 (0.013)	0.153 (0.019)
f_4	0.115 (0.023)	0.117 (0.011)	0.118 (0.015)	0.118 (0.022)	0.096 (0.009)	0.123 (0.013)
f_5	0.094 (0.030)	0.098 (0.013)	0.098 (0.019)	0.096 (0.025)	0.080 (0.012)	0.103 (0.017)
f_6	0.036 (0.016)	0.034 (0.007)	0.034 (0.009)	0.037 (0.022)	0.028 (0.007)	0.037 (0.010)
f_7	0.032 (0.013)	0.033 (0.006)	0.032 (0.008)	0.033 (0.013)	0.026 (0.006)	0.034 (0.009)

[a] The mean (standard deviation) of $\hat{\mathbf{f}}$ obtained from the weighted EM algorithm using both cases and controls.
[b] The mean (standard deviation) of $\hat{\mathbf{f}}$ obtained from the ordinary EM algorithm using both cases and controls.
[c] The mean (standard deviation) of $\hat{\mathbf{f}}$ obtained from the ordinary EM algorithm using only controls.

The performance of the proposed methods for estimating relative-risk parameters (β) is shown in Table 2. We observe that average estimates of the relative-risk parameters ($\bar{\hat{\beta}}$) were very close to the true parameter values for all of the methods. A comparison of the empirical SD of the estimates (SD_E) with and without known phase information ("Cox" and "Exact") showed that phase ambiguity could result in a significant loss of precision. The average estimates of the SD under the rare-disease approximation ($\overline{SD}_A$) appeared to be very close to the SD_E. The corresponding simulated coverage probabilities ("95% coverage") were also generally close to the nominal level of 95%.

We evaluated the consistency of the proposed estimator for the cumulative baseline hazard function under the nested case–control design (equation (6)). For each simulated data set, we fit a linear regression model, without the intercept term, to the estimates of $\Lambda_0(t)$ at the observed event times. By comparing the corresponding averaged regression coefficients to the true value of λ_0, we found that the proposed estimator for the cumulative baseline hazard function performed well. In the simulation setting of Table 2, for example, the true λ_0 was 0.182. The $\hat{\lambda}_0$ were 0.197 and 0.190 for $\beta = (0, 0, 0)$ and $\beta = (1.5, 0.4, 0)$, respectively.

6.2 *Efficiency of the Proposed Method for Cohort Studies Relative to the NPMLE Approach of Lin* (2004)

We conducted simulation studies to compare the proposed method with the NPMLE approach of Lin (2004). We considered a simple scenario involving two di-allelic loci with four haplotypes (AB, Ab, aB, aa), where A/a and B/b are the two alleles at the first and the second locus, respectively. We assumed all of the four haplotypes are equally frequent ($\mathbf{f}$ = 0.25), a setting that guarantees a substantial amount of phase ambiguity in the corresponding genotype data. We generated diplotype data for the cohort of subjects as before. We generated time-to-disease onset (T) from a Weibull distribution assuming that the haplotype Ab is associated with the risk of disease with a corresponding log-relative-risk (β) parameter

Table 2

*Simulation study involving GPX*1 *gene: performance of different estimators of relative-risk parameters. Cohort size* = 3000, *control/case matching ratio* = 1, *number of cases* ≈ 300.

				Rare disease		
β	True value	Cox (SD_E)[a]	Exact (SD_E)[b]	Rare (SD_E)[c]	$\overline{SD}_A$[d]	95% coverage[e]
			Under the null			
β_1	0	−0.006 (0.139)	−0.012 (0.190)	−0.012 (0.190)	0.201	97.8%
β_2	0	−0.023 (0.181)	−0.035 (0.237)	−0.034 (0.242)	0.247	95.6%
β_3	0	−0.008 (0.144)	−0.008 (0.158)	−0.008 (0.158)	0.168	97.3%
			Under the alternative			
β_1	1.5	1.526 (0.198)	1.534 (0.248)	1.499 (0.219)	0.220	93.8%
β_2	0.4	0.427 (0.242)	0.420 (0.306)	0.422 (0.286)	0.268	90.8%
β_3	0	0.008 (0.193)	0.005 (0.207)	0.012 (0.204)	0.202	96.4%

[a] The mean (standard deviation) of $\hat{\beta}$ obtained using standard Cox analysis assuming known phase.
[b] The mean (standard deviation) of $\hat{\beta}$ obtained using the proposed method without the rare-disease approximation, assuming unknown phase information.
[c] The mean (standard deviation) of $\hat{\beta}$ obtained using the proposed method with the rare-disease approximation, assuming unknown phase information.
[d] The mean of estimated standard deviations obtained from the asymptotic variance–covariance formula under the rare-disease approximation.
[e] 95% coverage probabilities.

Table 3

Cohort studies: comparison of the proposed method for estimating β and $\mathbf{f}$ *with the NPMLE procedure of Lin* (2004)*. Cohort size* = 1000. $\mathbf{f} = (0.25, 0.25, 0.25, 0.25)$.

		350 cases			100 cases		
True β		NPMLE[a]	Exact[b]	Rare disease[c]	NPMLE	Exact	Rare disease
1.5	$\bar{\hat{\beta}}$(SD)	1.498 (0.140)	1.499 (0.143)	1.360 (0.121)	1.509 (0.256)	1.516 (0.255)	1.480 (0.243)
	$\bar{\hat{\mathbf{f}}}_1$(SD)	0.250 (0.011)	0.250 (0.010)[d]		0.250 (0.012)	0.250 (0.012)	
	$\bar{\hat{\mathbf{f}}}_2$(SD)	0.249 (0.011)	0.250 (0.010)		0.250 (0.012)	0.250 (0.011)	
	$\bar{\hat{\mathbf{f}}}_3$(SD)	0.251 (0.011)	0.251 (0.011)		0.250 (0.012)	0.250 (0.011)	
	$\bar{\hat{\mathbf{f}}}_4$(SD)	0.250 (0.011)	0.250 (0.010)		0.250 (0.012)	0.250 (0.012)	
0	$\bar{\hat{\beta}}$(SD)	0.001 (0.137)	0.001 (0.137)	0.000 (0.137)	0.031 (0.247)	0.031 (0.248)	0.032 (0.248)
	$\bar{\hat{\mathbf{f}}}_1$(SD)	0.249 (0.011)	0.249 (0.011)		0.250 (0.012)	0.250 (0.011)	
	$\bar{\hat{\mathbf{f}}}_2$(SD)	0.250 (0.011)	0.250 (0.011)		0.250 (0.011)	0.250 (0.011)	
	$\bar{\hat{\mathbf{f}}}_3$(SD)	0.251 (0.012)	0.251 (0.011)		0.250 (0.011)	0.250 (0.011)	
	$\bar{\hat{\mathbf{f}}}_4$(SD)	0.250 (0.012)	0.250 (0.011)		0.249 (0.011)	0.249 (0.010)	

[a] The NPMLE (average and empirical standard deviation) of Lin (2004).

[b] The proposed method without the rare-disease assumption (average and empirical standard deviation).

[c] The proposed method with the rare-disease assumption (average and empirical standard deviation).

[d] The proposed method adopted the EM algorithm of Excoffier and Slatkin (1995) based on the full cohort.

of either 0 or 1.5. The shape and the scale parameters for the Weibull distribution were chosen in such a way that the overall disease rate in the population was on average 35% or 10%.

For estimation of the relative-risk parameters (Table 3), the rare-disease approximation for the proposed method led to negligible bias except when the disease rate was as high as 35% and the true value of β was 1.5. The bias in the latter situation was noticeable, but modest. The SDs for the proposed method, with or without the rare-disease assumption, were very close to those for NPMLE in all of the scenarios considered. The bias and SDs of the estimates of haplotype frequencies using the ordinary EM algorithm applied to the whole cohort were very similar to those obtained from the NPMLE method (Table 3). When the disease rate was 10%, which resulted in a small number of cases, the proposed algorithm for estimating the haplotype frequencies and relative-risk parameters in separate steps seemed to have better convergence properties than the NPMLE method that jointly estimates the two sets of parameters. In particular, the NPMLE method failed to converge within 500 iterations for 17 out of the 200 simulations.

7. Discussion

A comparison of the proposed method with the NPMLE procedure proposed by Lin (2004) for cohort studies merits further discussion. A major advantage of the proposed method is that it is applicable not only to cohort studies but also to nested case–control studies. Full-cohort studies, although popularly used for common traits such as heart disease, are not practical for the study of rare diseases such as cancer, as the study may require genotyping and expensive ascertainment of environmental exposures for an unnecessarily large number of subjects. Thus, many existing cohorts, such as the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute, are now being used for conducting nested case–control studies of genetic susceptibilities and gene–environment interactions.

The NPMLE method of Lin (2004) estimates haplotype frequency parameters ($\mathbf{f}$) jointly with the parameters β and $\lambda_0(t)$ of the CPH model. In contrast, we propose to estimate the haplotype frequencies ($\mathbf{f}$) in a separate step, completely independent of the estimation of β and $\lambda_0(t)$. For both cohort studies and nested case–control studies, assuming the rare-disease approximation for the latter design, we propose estimating $\mathbf{f}$ based on an ordinary EM algorithm that has been widely used for estimating haplotype frequencies from unphased genotype data. This allows one to take advantage of the existing computationally efficient programs (e.g., Niu et al., 2002), which could be particularly useful when a large number of SNPs are involved. In our limited simulation studies, where we compared the performance of the proposed method with that of NPMLE, we also found that estimating the haplotype frequencies independently of parameters of the CPH model led to a more stable and faster algorithm of parameter estimation.

Under the rare-disease approximation, the standard form of the proposed $\mathrm{PL}^{\mathrm{rare}}$ leads to several practical advantages. When ties are present, standard solutions for cohort and nested case–control studies (Breslow, 1974; Borgan and Langholz, 1993) can be applied. Modification of our approach can also be easily developed for other study designs, such as the counter-matching design (Langholz and Borgan, 1995) and the case–cohort design (Prentice, 1986). For nested case–control studies, matching on several time-dependent factors, such as age and calendar year, can also be handled without additional complications.

The NPMLE procedure of Lin (2004) is asymptotically most efficient under the setting of cohort design. We conducted limited simulation studies to compare the bias and efficiency of the proposed method, with or without the rare-disease approximation, with those of NPMLE. We found that the rare-disease approximation for the proposed method led to quite small bias even when the disease was relatively common with an overall prevalence of 35%. The loss of efficiency of the proposed method compared to the NPMLE was also negligible or minimal. These results are encouraging given that the proposed method, with the rare-disease assumption, is computationally very simple and can be easily generalized to various alternative designs. In future, however, more elaborate simulation studies are needed to compare the two methods in a wider variety of situations.

Acknowledgements

The authors thank the ATBC Study investigators at the National Cancer Institute for allowing us to use their data to demonstrate our methods. The authors also thank the editors and reviewers for their helpful comments.

References

Borgan, O. and Langholz, B. (1993). Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics* **49,** 593–602.

Borgan, O., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23,** 1749–1778.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* **33,** 228–237.

Breslow, N. E. (1972). Contribution to the discussion on the paper by DR Cox, Regression and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 216–217.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30,** 89–99.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **34,** 187–220.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73,** 1316–1329.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12,** 921–927.

Fallin, D. and Schork, N. (2000). Accuracy of haplotype frequency estimation for biallelic loci via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics* **67,** 947–959.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72,** 147–148.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics* **20,** 1903–1928.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Langholz, B. and Borgan, O. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika* **82,** 69–79.

Lin, D. Y. (2004). Haplotype-based association analysis in cohort studies of unrelated individuals. *Genetic Epidemiology* **26,** 255–264.

Niu, T., Qin, Z., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *American Journal of Human Genetics* **70,** 157–169.

Oakes, D. (1981). Survival times: Aspects of partial likelihood. *International Statistical Review* **49,** 235–264.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84,** 379–394.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70,** 425–434.

Stram, D., Pearce, C. L., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., and Thomas, D. C. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55,** 179–190.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining, by F. D. K. Liddell, J. C. McDonald, and D. C. Thomas. *Journal of the Royal Statistical Society, Series A* **140,** 469–491.

Wallenstein, S., Hodge, S., and Weston, A. (1998). A logistic regression model for analyzing extended haplotype data. *Genetic Epidemiology* **15,** 173–181.

Woodson, K., Tangrea, J. A., Lehman, T. A., Modali, R., Taylor, K. M., Snyder, K., Taylor, P. R., Virtamo, J., and Albanes, D. (2003). Manganese superoxide dismutase (MNSOD) polymorphism, alpha-tocopherol supplementation and prostate cancer risk in the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (Finland). *Cancer Causes and Control* **14,** 513–518.

Zhao, L. P., Li, S., and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics* **72,** 1231–1250.

Received April 2004. *Revised April* 2005.
Accepted April 2005.

Appendix

We derive the asymptotic variance formula for the proposed method under the rare-disease approximation. Here we assume a nested case–control design but note that the same derivation follows for the cohort design with only slight change of notation.

A.1 *Consistency*

We first show the approximate unbiasedness of $U(\beta,\ \mathbf{f})$. The proof and notation here closely follow results in Borgan and Langholz (1993) and Borgan, Goldstein, and Langholz (1995). Define $\mathcal{F}_{t-}$ to be the filtration containing information on (Δ, X) in the cohort, genotype $\mathbf{G}$, and the sampling information in the interval $[0,\ t)$. Define $N_{(i,r)}(t)$ to be the counting process for the observed number of failures for the ith subject in $[0,\ t]$ with associated sampled risk set r. Then $N_{(i,r)}(t)$ has intensity $\lambda_{(i,r)}(t) = \lambda_0(t)e^{\beta D_i}\pi_t(r \mid i)$, where $\pi_t(r \mid i)$ is the conditional probability that subset r is chosen from the risk set at event time t given $\mathcal{F}_{t-}$ and that i fails at t. Then $M_{(i,r)}(t) = N_{(i,r)}(t) - \int_0^t \lambda_{(i,r)}(u)\,du$ is a local square integrable martingale with respect to filtration $\mathcal{F}_{t-}$. The random variables $M_{(i,r)}(t)$ and $M_{(i,r)}(s)$ at different time points t and s are uncorrelated and have mean zero.

Let $s(D)$ be the design vector corresponding to the diplotype D, and let $\mathcal{P}^m$ denote collection of all subsets of $\{1, 2, \ldots, n\}$ of size m. For any function of diplotype D, $c(D)$, let $\mathrm{E}[c(D) \mid \mathbf{G}] = \sum_{D \in \mathcal{D}_{\mathbf{G}}} c(D)\mathrm{pr}_{\mathbf{f}}(D \mid \mathbf{G})$. Without loss of generality, we assume that the study is conducted in a fixed time period τ. The score function for β can be written as

$$U(\beta, \mathbf{f}; \tau)$$

$$= \frac{1}{n}\int_0^\tau \sum_{r \in \mathcal{P}^m} \sum_{i \in r} \left[\frac{\mathrm{E}\left[s(D)e^{\beta s(D)} \,\middle|\, \mathbf{G}_i\right]}{\mathrm{E}\left[e^{\beta s(D)} \,\middle|\, \mathbf{G}_i\right]} - \frac{\sum_{l \in r} \mathrm{E}\left[s(D)e^{\beta s(D)} \,\middle|\, \mathbf{G}_l\right]}{\sum_{l \in r} \mathrm{E}\left[e^{\beta s(D)} \,\middle|\, \mathbf{G}_l\right]} \right] dM_{(i,r)}(t)$$

(Borgan et al., 1995). The integrand in the above formula is unrelated to time t and thus is predictable. By the standard martingale theory, the process $U(\beta,\ \mathbf{f};\ t)$ is a martingale so that $\mathrm{E}[U(\beta,\ \mathbf{f};\ \tau)] = 0$. Moreover, the estimate of $\mathbf{f}$ using data from controls only is approximately consistent under the rare-disease assumption. The existence and consistency of the estimate of β that solves $U(\beta, \hat{\mathbf{f}}; \tau) = 0$ now follow from the results given in Foutz (1977).

A.2 *Asymptotic Normality*

A standard Taylor's series expansion of $U(\hat{\beta}, \hat{\mathbf{f}})$ around the true parameter values $(\beta, \mathbf{f})$ leads to $n^{1/2}(\hat{\beta} - \beta) = I_{\beta\beta}^{-1}n^{1/2}U(\beta, \mathbf{f}; \tau) - I_{\beta\beta}^{-1}I_{\beta\mathbf{f}}n^{1/2}(\hat{\mathbf{f}} - \mathbf{f}) + o_p(1)$, where $I_{\beta\beta}$ and $I_{\beta\mathbf{f}}$ are the large-sample limits of $-\partial U(\beta,\ \mathbf{f})/\partial\beta$ and $-\partial U(\beta,\ \mathbf{f})/\partial\mathbf{f}$, respectively. Following the standard asymptotic theory for nested case–control studies (Borgan and Langholz, 1993), we have $\mathrm{cov}[U(\beta,\ \mathbf{f};\ \tau)] = I_{\beta\beta}$ and $n^{1/2}U(\beta, \mathbf{f}; \tau) \sim \mathrm{Normal}(0, I_{\beta\beta})$. Moreover, from standard parametric maximum-likelihood inference theory, we have $(n_c)^{1/2}(\hat{\mathbf{f}} - \mathbf{f}) \sim \mathrm{Normal}[0, (I_{\mathbf{ff}}^c)^{-1}]$, where n_c is the total number of nonreplicated controls and $I_{\mathbf{ff}}^c$ is the asymptotic information matrix for $\mathbf{f}$ (Excoffier and Slatkin, 1995).

Furthermore, $U(\beta,\ \mathbf{f};\ \tau)$ and $(\hat{\mathbf{f}} - \mathbf{f})$ are asymptotically uncorrelated, which can be shown as follows. Let $\dot{l}^p(\mathbf{G}_j; \mathbf{f})$ be the jth control's contribution to the maximum-likelihood score function for $\mathbf{f}$. Then

$$\sqrt{n_c}(\hat{\mathbf{f}} - \mathbf{f}) = \frac{1}{\sqrt{n_c}} \sum_{j=1}^{n_c} \dot{l}^p(\mathbf{G}_j; \mathbf{f}) + o_p(1).$$

Let $H(\beta, \mathbf{f}; t) = U(\beta, \mathbf{f}; t)\sum_{j=1}^{n_c} \dot{l}^p(\mathbf{G}_j; \mathbf{f})$. We need to prove $\mathrm{E}H(\beta,\ \mathbf{f};\ t) = 0$. The proof follows by verifying that $H(\beta,\ \mathbf{f};\ t)$ is a martingale or equivalently the condition $\mathrm{E}[dH(\beta, \mathbf{f}; t) \mid \mathcal{F}_{t-}] = 0$. To show this, we note that $\mathrm{E}[dH(\beta, \mathbf{f}; t) \mid \mathcal{F}_{t-}] = \mathrm{E}[\sum_{j=1}^{n_c} \dot{l}^p(\mathbf{G}_j; \mathbf{f})\, dU(\beta, \mathbf{f}; t) \mid \mathcal{F}_{t-}]$. Since $\mathcal{F}_{t-}$ contains genotype information $\mathbf{G}$ for all subjects in the full cohort, we have $\mathrm{E}[dH(\beta, \mathbf{f}; t) \mid \mathcal{F}_{t-}] = \mathrm{E}\{dU(\beta, \mathbf{f}; t) \mid \mathcal{F}_{t-}\} \sum_{j=1}^{n_c} \dot{l}^p(\mathbf{G}_j; \mathbf{f})$, which is equal to zero as $U(\beta,\ \mathbf{f};\ t)$ is a martingale.

We further assume that as n goes to infinity, n/n_c converges to a fixed number ρ. For one-to-one matching, for example, ρ could be roughly equal to the disease prevalence in the cohort. The results above show that $n^{1/2}(\hat{\beta} - \beta)$ follows an asymptotically normal distribution with variance $\Sigma = I_{\beta\beta}^{-1} + \rho I_{\beta\beta}^{-1} I_{\beta\mathrm{f}}(I_{\mathbf{ff}}^{n_c})^{-1} I_{\beta\mathrm{f}}^T I_{\beta\beta}^{-1}$. Here Σ can be estimated as follows. Let $\hat{I}_{\beta\beta} = -\partial U(\beta, \mathbf{f})/\partial\beta|_{\hat{\beta},\hat{\mathbf{f}}}$, $\hat{I}_{\beta\mathbf{f}} = -\partial U(\beta, \mathbf{f})/\partial\mathbf{f}|_{\hat{\beta},\hat{\mathbf{f}}}$, and $\hat{I}_{\mathbf{ff}}^{n_c}$ be the estimated information matrix for $\mathbf{f}$ using controls only. Then Σ can be consistently estimated as $\hat{\Sigma} = \hat{I}_{\beta\beta}^{-1} + \hat{I}_{\beta\beta}^{-1}\hat{I}_{\beta\mathbf{f}}(\hat{I}_{\mathbf{ff}}^{n_c})^{-1}\hat{I}_{\beta\mathbf{f}}^T\hat{I}_{\beta\beta}^{-1}n/n_c$. Above, $\hat{I}_{\beta\mathbf{f}}$ and $\hat{I}_{\beta\beta}$ can be easily obtained by suitable analytical or numerical derivatives of $U(\hat{\beta}, \hat{\mathbf{f}})$.